

LA-UR -82-1893

CONF-821021--1

LA-UR--82-1893

DEG2 019576

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36.

MASTER


TITLE: ASPECTS OF MODEL SELECTION IN MULTIVARIATE ANALYSES

AUTHOR(S): Richard Picard, S-1

SUBMITTED TO: 1982 DOE Statistical Symposium, Idaho Falls, Idaho, October

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MP

By acceptance of this article, the publisher  nizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

ASPECTS OF MODEL SELECTION IN MULTIVARIATE ANALYSES

1. Introduction

Issues of model selection arise in many statistical problems that deal with a large number of observed variables. Some of the more commonly encountered problems involve use of multiple regression techniques, where it is often desired to find a relatively simple function that models some underlying phenomenon. Many statistical package programs (BMDP, SAS, SPSS, etc.) contain a number of subset selection algorithms (e.g., forward and backwards stepwise methods) for this purpose. There is a good deal of literature on the subject of choosing "the" algorithm that will produce a "best" model, but much less has been done concerning proper interpretation of the selected fitted equation. This is especially true in regard to assessment of the predictive ability of the model chosen.

Similar situations accompany other types of problems. Discriminant analysis is in many ways like multiple regression: a goal is to obtain a discriminant function to serve as a basis for classification of "future" objects into one of a number of groups. The choice of a specific discriminant function to be used and estimation of its associated misclassification probabilities are issues that resemble model selection and assessment in multiple regression.

Still other types of statistical problems entail reduction of a large number of variables to a more manageable collection. One example of this is the study of sensitivity analysis for reactor simulation codes. In some cases, there are hundreds of input variables of interest, and development of an understandable explanation of the system requires model building with an eye towards limiting the dimension of the input space.

The purpose of this paper is to examine aspects of variable selection procedures, particularly in terms of interpreting the result obtained. For clarity of the presentation, the discussion is pursued in the context of multiple regression.

i. The Optimism Principle

A major tenet of conventional statistical folklore is that a model chosen via some selection process provides a much more "optimistic" explanation of the data used in its derivation than it does of other data that will arise in a similar fashion. One of the more eloquent statements of this principle is

"Testing the procedure on the data that gave it birth is almost certain to overestimate performance, for the optimizing process that chose it from among many possible procedures will have made the greatest use possible of any and all idiosyncrasies of those particular data... As a result, the procedure will likely work better for these data than for almost any other data that will arise in practice." Mosteller and Tukey (1977), p. 37.

This doctrine appears to be based on a long history of unfortunate experiences encountered by statisticians.

The most important words in the above quotation are "from among many possible procedures," as the selection process plays a key role. Providing a formal demonstration of this phenomenon for subset selection procedures in multiple regression is not difficult. Consider the general linear model

$$\underline{y} = \underline{X}\underline{\beta} + \underline{e} \quad (2.1)$$

where

\underline{X} is an $n \times p$ matrix of constants of rank $p < n$,
 $\underline{\beta}$ is a p -component vector of unknown parameters,
and $\underline{e} \sim (0, \sigma^2 \underline{I})$ for σ^2 unknown.

The model (2.1) is called the "full model," and the possibility that some of the components of β are zero is often entertained. The purpose of subset selection algorithms is to extract a parsimonious fitted equation.

It is well known that the residual mean square from the least squares fit to the full model,

$$\hat{\sigma}_{full}^2 = \underline{y}' [I - X(X'X)^{-1} X'] \underline{y} / (n-p)$$

is an unbiased estimator of σ^2 . The same does not hold for residual mean squares from fitted models chosen by subset selection procedures. Suppose that all $2^p - 1$ least squares subset fits are considered and the "best" one is selected. If "best" is taken to mean the fit whose residual mean square is a minimum of those observed, then

$$\hat{\sigma}_{min}^2 \leq \hat{\sigma}_{full}^2 \quad \text{for all values of } y \quad (2.2)$$

where $\hat{\sigma}_{min}^2$ denotes the minimized value corresponding to the "best" fit. Since $\hat{\sigma}_{full}^2$ is unbiased for σ^2 over the distribution of y , it is obvious from (2.2) that $\hat{\sigma}_{min}^2$ is not. Clearly $\hat{\sigma}_{min}^2$ is "optimistic" and in some cases the bias can be substantial.

Here σ^2 represents a baseline measure of predictive ability: it is the minimum obtainable squared error of prediction of future values of y and corresponds to use of the "true" vector β in a predictor. That $\hat{\sigma}_{min}^2$ is on average less than σ^2 means that a naive user of the selected model could easily be misled into believing that he can predict much better than he actually will (and in some cases, better than is even possible).

A crude analogy can be drawn with the theory of order statistics. Suppose $(Z_i; i=1,2,\dots,m)$ is a random sample drawn from some distribution function $G(z)$. If $Z_{(1)}$ is the first order statistic, its distributional properties can be evaluated. For example,

$$\Pr(Z_{(1)} > c) = \left[\int_c^\infty dG(z) \right]^m$$

In short, the distribution function of $Z_{(1)}$ takes into account the fact that $Z_{(1)}$ is the smallest of the observed $\{Z_i\}$. In multiple regression, if a particular subset model is selected because it minimizes the observed value of some criterion - such as a residual mean square, C_p statistic (Mallows, 1973), or PRESS statistic (Allen, 1971) - the distributional properties of the observed optimized criterion must formally account for the selection process.

Unlike the theory of order statistics, however, exact distributional results are nearly impossible to obtain for model selection in the multiple regression framework. Nonetheless, simulation work has been pursued and Berk (1978a) has found realistic examples where bias in residual mean squares from fitted models chosen via stepwise algorithms exceeded 20% of the actual value of σ^2 . Many other statistics are similarly biased over the subset selection; observed values of C_p and PRESS tend to be much lower than might otherwise be expected (Berk, 1978b) while values of R^2 are greatly inflated (Diehr and Hoflin, 1974; Rencher and Pun, 1980).

The magnitudes of the biases in the usual summary statistics depend on a number of factors. The problems seem to be most serious when:

1. The number of observations, n , is of moderate size (say, less than 50) and the number of independent variables, p , is appreciable. For many variable selection problems, p of 10 or 20 is not out of the ordinary so that "best" fit may be chosen from literally thousands ($2^p - 1$) of candidates. When this degree of choice is available, substantial optimism can be induced.
2. The number of observations is small. Here a good deal of variability exists in statistics such as residual mean squares, and selecting the "best" from such a class of items can lead to optimism as well.

The main point to keep in mind is that imitation of standard statistical techniques that assume the form of the model is "known" - as opposed to selected - can lead to conclusions very much in error. For example, substitution of $\hat{\sigma}_{\min}^2$ into the usual formulas for confidence intervals is not justified theoretically and ensuing optimism can lead to a number of problems. In some sense, application (well, abuse) of the standard methodology has failed and alternatives must be considered.

3. Predictive Ability of Regression Models

It is not difficult to obtain good assessments of the predictive ability of fitted models that are the product of subset selection procedures. In order to develop the ideas behind such assessments, it is necessary to first introduce some notation.

Similar to the previous section, denote the general linear model

$$\underline{y} = \underline{X}\underline{\beta} + \underline{e}$$

where \underline{X} , $\underline{\beta}$, and \underline{e} are as in (2.1). Consider any subset selection method, such as a forward stepwise algorithm. The method has the property that given an observed \underline{y}_0 , a p -component vector $\underline{\beta}_0$ is produced for purposes of parameter estimation. If the full model is not selected, some of the components of $\underline{\beta}_0$ will be zero. The distribution of \underline{y} together with the selection algorithm thus induces a distribution on R^p for the ensuing estimator $\hat{\underline{\beta}}$. Assume the induced distribution is well enough behaved so that moments of order two exist and it is possible to write

$$\hat{\underline{\beta}} \sim (\underline{\mu}_{\hat{\underline{\beta}}}, \underline{\Sigma}_{\hat{\underline{\beta}}}) \quad (3.1)$$

Given this characterization of estimators produced by subset selection, the predictive ability of the associated fitted model can be evaluated. Suppose \underline{x}_f is a known p -component vector and

$$y_f = \underline{x}_f' \underline{\beta} + e_f \quad \text{for} \quad e_f \sim (0, \sigma^2)$$

is an observation conforming to the structure of (2.1) and independent of $\hat{\underline{\beta}}$. If the realized (y, X) is the product of some physical mechanism that has generated the data, (y_f, \underline{x}_f') can be thought of as a "future observation" from the same mechanism. The predicted value of y_f based on the selected model is $\hat{y}_f = \underline{x}_f' \hat{\underline{\beta}}$ and the error of prediction is $y_f - \hat{y}_f$.

The predictive ability of the selected model is reflected by the distribution of $y_f - \hat{y}_f$ for different choices of \underline{x}_f . A useful criterion for evaluating prediction is the mean squared error

$$\begin{aligned} \text{MSE}(\underline{x}_f) &= E(y_f - \hat{y}_f)^2 \\ &= \sigma^2 + (\underline{\beta} - \underline{\mu}_{\hat{\beta}})' \underline{x}_f \underline{x}_f' (\underline{\beta} - \underline{\mu}_{\hat{\beta}}) + \text{tr } \hat{\Sigma}_{\hat{\beta}} \underline{x}_f \underline{x}_f' \end{aligned}$$

An overall measure of predictive ability can be obtained by integrating the mean squared error of prediction at \underline{x}_f with respect to a distribution F on R^p for \underline{x}_f . The integral

$$\int_{R^p} \underline{x} \underline{x}' dF = C$$

where C is the matrix of expected cross products defines

$$\begin{aligned} \text{IMSE}_F(\underline{\beta}) &= \int_{R^p} \text{MSE}(\underline{x}_f) dF \\ &= \sigma^2 + (\underline{\beta} - \underline{\mu}_{\hat{\beta}})' C (\underline{\beta} - \underline{\mu}_{\hat{\beta}}) + \text{tr } \hat{\Sigma}_{\hat{\beta}} C \end{aligned} \quad (3.2)$$

Here F may be taken to be any distribution on R^p , so that $\text{IMSE}_F(\underline{\beta})$ loosely denotes the predictive ability of the fitted model over the region of the design space "highlighted" by F . Of course, other types of a "loss function" besides mean squared error could be considered and a similar development pursued.

Expression (3.2) reflects two important aspects of prediction. The first is that a "diminishing returns" effect sets in, and increasing the number of observations used to derive $\hat{\beta}$ past a certain point does very little to improve predictive ability. This can be easily demonstrated in two simple examples.

Example: Suppose $\{u_i\}$, $i=1,2,\dots,n$, are iid $N(\mu, \sigma^2)$ and it is of interest to predict a "future" observation u from the same population. A common predictor for this case is $\bar{u}_n = \sum u_i/n$, whose standard error of prediction is

$$\sqrt{E(u - \bar{u}_n)^2} = \sigma \sqrt{1 + (1/n)}$$

For $n = 10$ this standard error is 1.049σ , while for $n = 100$ it is 1.005σ and declines only to 1.000σ as n approaches infinity. The percentage reduction in the standard error due to increasing n beyond 10 is quite small, and reflects that the component of error due to prediction of the future very quickly dominates the component of error due to estimation of the parameter μ .

Example: In the general linear model (2.1), partition $X = [X_1 \ X_2]$ and $\beta' = [\beta_1' \ \beta_2']$ so that the model can be written

$$y = X_1 \beta_1 + X_2 \beta_2 + e$$

Consider the estimator

$$\hat{\beta}_u = \begin{pmatrix} (X_1' X_1)^{-1} X_1' y \\ 0 \end{pmatrix}$$

obtained by "underfitting" the full model using ordinary least squares. Moments of $\hat{\beta}_u$ are well known and substitution into (3.2) gives

$$IMSE_F(\hat{\beta}_u) = \sigma^2 + \beta_2' [-A_X' \ I] C \begin{bmatrix} -A_X \\ I \end{bmatrix} \beta_2 + \sigma^2 \text{tr} C \begin{pmatrix} (X_1' X_1)^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

where $A_X = (X_1'X_1)^{-1}X_1'X_2$ is the alias matrix for the design. The three terms in $\text{IMSE}_F(\hat{\beta}_u)$ have simple interpretations. The first, σ^2 , can be thought of as error due solely to prediction of the "future" and can never be reduced by taking more observations. The second term,

$$\beta_2' [-A_X' \ I] C \begin{bmatrix} -A_X \\ 1 \end{bmatrix} \beta_2 = \int_{R^p} [x'(\beta - E\hat{\beta}_u)]^2 dF$$

is an average squared bias and, when $\beta_2 \neq 0$, represents a penalty for fitting a model of the wrong form. This term also does not decrease intrinsically with sample size: for a "doubled" data set of $2n$ observations,

$$X^* = \begin{pmatrix} X \\ X \end{pmatrix} \text{ implies } A_{X^*} = A_X$$

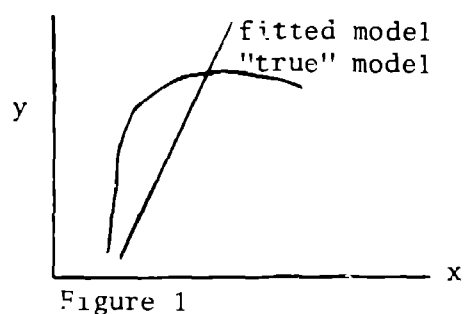
and bias properties remain identical to those for the fit to n observations. The third term of $\text{IMSE}_F(\hat{\beta}_u)$,

$$\sigma^2 \text{tr} C \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \int_{R^p} \text{Var}(x'\hat{\beta}_u) dF$$

is the component of error due to the variability in parameter estimation. This term tends to decline as $1/n$ and in most cases is dominated by the sum of the first two terms. It follows that $\text{IMSE}_F(\hat{\beta}_u)$ is not particularly sensitive to the magnitude of n .

As fitted models that are the product of subset selection procedures are quite difficult to evaluate theoretically, a rigorous demonstration of diminishing returns for them along the lines of the above examples is not easy to come by. However, there seems little reason to believe that the predictive ability of such models behaves in a fundamentally different fashion in this regard.

The second important aspect of prediction using regression fits is the dependence of $IMSE_F(\hat{\beta})$ on F . This phenomenon is apparent from inspection of (3.2) and, crudely stated, means that fitted models predict better at some values of x_f than at others. When the fitted model is of the wrong form, this is especially true. One simple illustration is provided in Figure 1, where the "true" model exhibits some curvature but the fitted model does not. Effects of outliers and heteroscedasticity on fitted models can also



exaggerate differences in predictive ability at different values of x_f . This "localized" nature of prediction should be kept in mind in analysis of many problems.

4. Analysis of Multivariate Data

Often in the analysis of large data sets, the "right" way to proceed is not immediately apparent. Consequently, some aspects of model selection or so-called exploratory data analysis inevitably arise. Competing models may be temporarily entertained and the final choice of a predictive equation (or discriminant function or whatever) is influenced by many factors, including the personal experiences and prejudices of the people involved. This state of affairs is not likely to change in the near future.

It is very important to develop the underpinnings of proper assessment of models produced by such selection procedures. The main point to keep in mind in this regard is the optimism principle: when a model is chosen because of qualities exhibited with respect to a particular set of data, its "explanation" of future observations that arise in a similar fashion will almost certainly not be as good as would naively be expected based on the original data. As illustrated in the specific examples of Section 2, blind application of ordinary statistical methods that fail to account for the selection process can lead to incorrect conclusions. Since it is usually impractical to attempt derivation of a theory to rigorously handle the optimism in such analyses, alternatives must be considered.

A very old concept can be adapted to deal with this problem. Basically, the approach splits the data into two portions (not necessarily of equal size) before analysis to obtain a fitted model. The selection can be applied to one portion alone, and assessment can be made using the other portion. For regression problems, the general strategy behind this approach is to take advantage of the diminishing returns effect to develop "almost" as good a predictor from a portion of the data as would have been obtained using the complete set and simultaneously avoid the optimism principle by basing the assessment on observations not used in the model selection. Ideas along these lines have been proposed since the 1930's - see Stone (1974) for a brief historical account.

Despite its early origins, the subject of cross-validation has not yet been thoroughly examined. Specific details for implementation of a "splitting" strategy have not been laid out in detail, particularly in regard to underlying theoretical justifications of the methods involved. Snee (1977) has suggested using the DUPLEX algorithm to determine the split and then taking the observed mean squared error of prediction of the "validation" portion as an overall evaluation of the fitted model. Mosteller and Tukey (1977) have proposed a three-way-split, using one portion of the data to divine the functional form of the model, a second portion to estimate parameters, and a third portion for assessment. Still others have simplistically suggested a purely random division of the data.

The principles behind splitting are not difficult to grasp. It is important to maintain the integrity of the validation sample, so as to allow the data to "play the role" of future observations produced by the physical mechanism of interest. In this regard, it is presumed that the stated full model is correct; otherwise it may not be possible to obtain a realistic evaluation of predictive ability. For example, in an experiment where substantial day to day variation exists but the observed data are collected from only one day, "actual variability" would likely be underestimated based on the observations at hand.

When splitting a data set, the size of the validation portion does not appear to be crucial (within broad limits) and primarily should reflect sentiments concerning the importance of developing a predictor relative to assessing what has been developed. As for which observations are placed into the respective portions, principles of sampling apply. Some have suggested a random division, but because of the localized nature of predictive ability an attempt should be made to assure that data from all regions of the design space are represented in both portions. When the number of variables is large, "gaps" can easily result in splits determined by a simple random sampling procedure. This motivates an approach similar in nature to stratified sampling, where the strata here loosely correspond to different regions of the design space. The DUPLEX algorithm incorporates this notion, and other methods also suggest themselves.

Using the validation sample (however chosen) is fairly straightforward. The main point to keep in mind is that predictive ability often differs depending on the region of interest in the design space. This is particularly true when the fitted model is not of the correct form. This motivates a "localized" use of validation residuals for assessment, though for most purposes an examination of plots of these residuals can be the best way to acquire an understanding of the qualities of the model.

5. Summary

Analysis of data sets that involve large numbers of variables usually entails some type of model fitting and data reduction. In regression problems, a fitted model that is obtained by a selection process can be difficult to evaluate because of optimism induced by the choice mechanism. Problems in areas such as discriminant analysis, calibration, and the like often lead to similar difficulties. The preceding sections reviewed some of the general ideas behind assessment of regression-type predictors and illustrated how they can be easily incorporated into a standard data analysis.

References

1. Allen, D. M. (1971). The Prediction Sum of Squares as a Criterion for Selecting Predictor Variables, Technical Report No. 23, Dept. of Statistics, Univ. of Kentucky.
2. Berk, K. (1978a). Comparing Subset Selection Procedures, Technometrics, 20, 1-6.
3. Berk, K. (1978b). Sequential PRESS, Forward Selection, and the Full Regression Model, ASA Proc. on Stat. Comp., 309-313.
4. Diehr, G. and Hoflin, D. R. (1974). Approximating the Distribution of the Sample R^2 in Best Subset Regressions, Technometrics, 16, 317-320.
5. Mallows, C. (1973). Some Comments on C_p , Technometrics, 15, 661-675.
6. Mosteller, F. and Tukey, J. W. (1977). Data Analysis and Regression, Addison-Wesley, Reading, Mass.
7. Rencher, A. C. and Pun, F. C. (1980). Inflation of R^2 in Best Subset Selection, Technometrics, 22, 49-53.
8. Snee, R. D. (1977). Validation of Regression Models: Methods and examples, Technometrics, 19, 415-428.
9. Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society, Series B, 36, 111-147.